# 基于 XLNet-TB 的中文文本可读性评估研究

倪佳成 [1]
NI Jiacheng

**摘 要** 文本可读性用于评估一段文本的阅读难度，这一概念在教育和出版领域发挥着重要作用。针对目前中文文本可读性评估模型在捕捉文本深层次语义信息方面的不足，文章提出了一种基于 XLNet-TB 分层网络架构的中文文本可读性评估模型，该模型利用 Chinese-XLNet 预训练模型生成文本词向量表示，然后联合使用 TextCNN 与 BiGRU 模型提取文本语义特征，最终输入全连接层获得文本可读性评估分级结果。实验证明，该模型在汉语水平考试 HSK 真题数据集上的准确率达到了 89.5%，验证了模型的有效性。

**关键词** 可读性评估；深度学习；XLNet 预训练模型；卷积神经网络；双向门控递归单元

doi：10.3969/j.issn.1672-9528.2025.04.014

## 0 引言

在学习语言、阅读文本时，选择与自身语言水平相近水平的文本可以更好地达到学习与阅读的目的。文本可读性是一项用来衡量文本难度的指标，它通过量化影响文本难度的多维度因素，来评估一篇文本的阅读难度，在教育与出版领域中有着重要的应用价值。目前，文本可读性的评估方法主要有 3 种：基于可读性公式、基于机器学习与基于深度学习的文本可读性评估方法。

可读性公式融合了文本阅读难度中诸多可量化的影响因素，使用公式的形式评估文本的难易程度。左虹等人 [1] 针对中级欧美留学生的汉语文本，提出了一种文本可读性计算公式，该公式融入字数、句子结构、词汇使用等多种语言特征。刘苗苗等人 [2] 利用多元线性回归模型，构建了一套汉语文本可读性计算公式。王蕾 [3] 通过实证研究提出了一种适用于中级日韩学习者的汉语文本可读性计算公式，主要指标有平均句子长度、词汇丰富度、难词比例等。可读性公式虽用法广泛且易于计算，但其往往只是通过选取几个指标来代表整个文本，难以捕捉文本深层次的内部语义特征，准确度较低。

随着机器学习技术的发展，机器学习的方法也被应用到文本可读性评估中。于东等人 [4] 通过对比支持向量机与逻辑回归两种机器学习算法在文本可读性计算中的表现，构建了基于逻辑回归算法的中文句子难度评估模型。杜月明等人 [5] 对比了 6 种机器学习模型的效果，从汉字、词汇、句法层面筛选出多个特征，构建了基于支持向量机算法的文本可读性评估模型。杨智渤 [6] 从汉字、词汇、句法层面提取 95 个特征，构建了基于随机森林的文本可读性分类器。颜伟嘉 [7] 提出了一种基于机器学习的文本可读性分析模型。首先利用显著性检验对可能影响文本可读性的特征进行筛选，之后对比多个机器学习模型，实验发现，支持向量机模型的效果最好。机器学习模型以其强大的学习能力与复杂的特征表示极大增强了文本可读性评估的准确度。但对于特征的选择，往往会耗费大量的时间与人工来进行复杂的人工设计，且不恰当的特征选择会导致模型性能的下降。

深度学习是一种更加高级的机器学习技术，对特征进行自动提取，大大减少了人工的参与。唐玉玲等人 [8] 使用深度学习模型对文本可读性进行评估，探究不同深度学习提取器的自动捕获文本特征的能力，验证了深度学习在文本可读性分类中的有效性，同时发现结果也优于普通的机器学习模型。Muhammad 等人 [9] 提出了一种基于 CNN 的高等教育文本可读性评估模型，在英文数据集中获得了良好的效果。沙政 [10] 针对英文文本，提出了基于混合网络模型的可读性评估模型，在英文文本数据集中取得了良好的效果。

上述研究展现出深度学习在文本可读性评估中的优秀性能，但现有模型仍在捕捉深层次语义信息方面存在一定的不足，从而导致可读性评估结果出现偏差。针对这一问题，本文提出了一种基于 XLNet-TB 分层网络架构的中文文本可读性评估模型，该模型利用 Chinese-XLNet 预训练模型生成文本的动态词向量表示，联合使用 TextCNN 与 BiGRU 模型自动提取文 本语义特征，强化模型捕捉深层次语义信息的能力，提高文本可读性评估的准确度。

## 1 XLNet-TB 模型概况

本文提出的 XLNet-TB 中文文本可读性评估模型的结构如图 1 所示。首先对原始文本数据进行预处理，其次利用

1. 南京审计大学计算机学院 江苏南京 211815

XLNet 层将中文文本转化为动态词向量，再经过 TextCNN 层提取文本局部语义信息特征，之后经过 BiGRU 层利用正向与反向 GRU 提取上下文全局语义信息特征，最终在全连接层进行分类并输出预测结果。

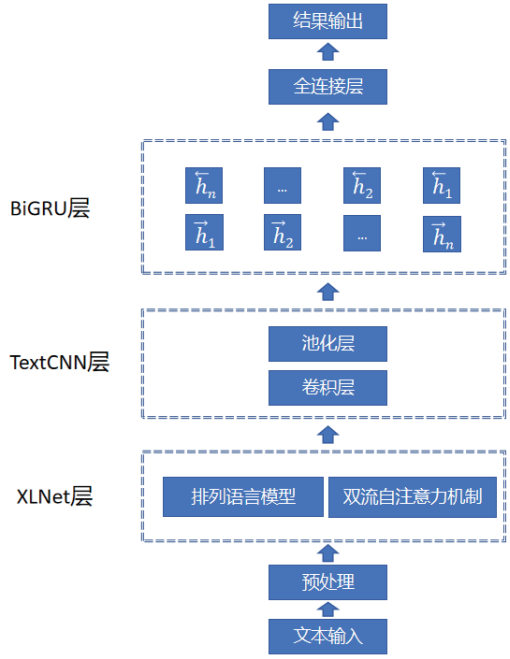


图 1 模型结构

### 1.1 XLNet 层

由于计算机无法直接处理文本语言，因此需要利用词向量预训练模型先将文本转化为词向量的形式。目前，词向量预训练模型主要分为自回归模型和自编码模型。自回归模型是一种单向的预测模型，通过上文或下文，预测出下一个会出现的字词；自编码模型可以利用上下文双向信息，对字词进行预测。通过对原句中加入掩码 Mask，破坏原句结构，再根据上下文预测出被遮掩的信息。

XLNet 是一种基于自回归的预训练语言模型。不同于一般的自回归模型，XLNet 在自回归模型的基础上引入了自编码模型的优点，改善了 BERT 被遮掩词之间无关联的问题，具体由排列语言模型与双流自注意力机制实现。

（1）排列语言模型

排列语言模型会打乱一句话的排列顺序，遮掩末尾若干个字词，最后以自回归的方式按照当前的排列顺序预测出被遮掩的字词。模型预测被遮掩词概率用公式表示为：

$$p_\theta(X_{z_t}=x|x_{z<t})=\frac{\exp[\boldsymbol{e}(x)^{\mathrm{T}}g_\theta(x_{z<t},z_t)]}{\sum_{x'}\exp[\boldsymbol{e}(x')^{\mathrm{T}}g_\theta(x_{z<t},z_t)]'} \quad (1)$$

式中：$X_{z_t}$为某排序中的第 $t$ 个元素；$x_{z<t}$为某排序中第一个元素到第 $t$-1 个元素；$g_\theta(x_{z<t},z_t)$为 Yang 等人[11]提出的模型，在给定位置 $z$ 之前的上下文$x_{z<t}$和当前输入 $z_t$ 时，用以捕捉上下文信息；$\boldsymbol{e}(x)$ 为 $x$ 的词向量；$x'$ 为当前预测词。

虽然可以对一个长度为 $T$ 的序列遍历出 $T!$ 种排列方式，让模型去学习所有的上下文语义信息，但是遍历 $T!$ 次，计算量过于庞大。所以 XLNet 只选择部分排序，通过计算序列所有排列方式的期望值并依此去选取合适的序列，计算公式为：

$$\max_\theta E_{z\sim Z_t}[\sum_{t=1}^{T}\log P_\theta(x_{z_t}|X_{z<t})] \quad (2)$$

式中：$z_t$ 表示长度为 $T$ 的序列中的全部排序所构成的集合；$z$ 表示其中一种排序方式；$E_{z\sim Z_t}$表示对所有的排序方式求期望。

（2）双流自注意力机制

排列语言模型虽然获取了上下文语义信息，但无法获取预测词的原始序列位置信息。这是由于传统自回归模型只能看到预测词之前的信息，不会打乱原始序列，而排列语言模型会重新打乱序列的位置。为获取预测词的原始序列位置信息以提升模型预测的准确率，XLNet 引入了双流自注意力机制。

双流自注意力机制旨在预测目标词时，既要得到目标词的位置信息，还要得到其他词的位置和内容信息。双流指内容流与查询流，内容流表示目标词内容信息，它能看到当前词的内容信息与位置信息。查询流表示原始输入序列中目标词的位置信息，只能看到当前词的位置信息，无法看到内容信息，内容流与查询流的计算公式分别为：

$$h_{z_t}^m \leftarrow \mathrm{Attention}(\boldsymbol{Q}=h_{z_t}^{m-1},\boldsymbol{KV}=h_{z\le t}^{m-1};\theta) \quad (3)$$

$$g_{z_t}^m \leftarrow \mathrm{Attention}(\boldsymbol{Q}=g_{z_t}^{m-1},\boldsymbol{KV}=h_{z<t}^{m-1};\theta) \quad (4)$$

式中：$h$ 为内容隐状态；$g$ 为查询隐状态；$m$ 为 XLNet 层数；$\boldsymbol{Q}$ 为查询向量；$\boldsymbol{K}$ 为待查询向量；$\boldsymbol{V}$ 为内容向量。

### 1.2 TextCNN 层

TextCNN 层由卷积层和池化层组成，能够有效地获取文本局部关键信息特征。

（1）卷积层

在卷积层，模型会将经 XLNet 层训练过的词向量矩阵（$b\times n\times d$）作为输入，$b$ 为句子数量，$n$ 为句子长度，$d$ 为词向量维度。随后，通过多个不同尺寸的卷积核对词向量矩阵进行卷积计算，以提取不同粒度的局部特征。特征 $C_i$ 的计算公式为：

$$C_i=\mathrm{ReLU}(\boldsymbol{W}_m\cdot\boldsymbol{X}_{i:i+h-1}+b) \quad (5)$$

式中：$W_m$ 表示多个不同体积的卷积核；$X_{i:i+h-1}$表示输入矩阵中第 $i$ 个词向量到第 $i$+$h$-1 个词向量；$b$ 表示偏差项。

（2）池化层

在池化层，由卷积层得到的不同尺度的特征向量会被压

缩至相同的长度，之后以最大池化的方式对其进行处理，得到最突出的特征，最后将若干个经池化过的向量拼接起来，传入 BiGRU 层中，计算公式为：

$$P_i = \max(C_i) \tag{6}$$

1.3 BiGRU 层

BiGRU 是一种循环神经网络模型，擅长处理序列数据，由一串正向 GRU 与一串反向 GRU 构成。BiGRU 的结构如图 2 所示。

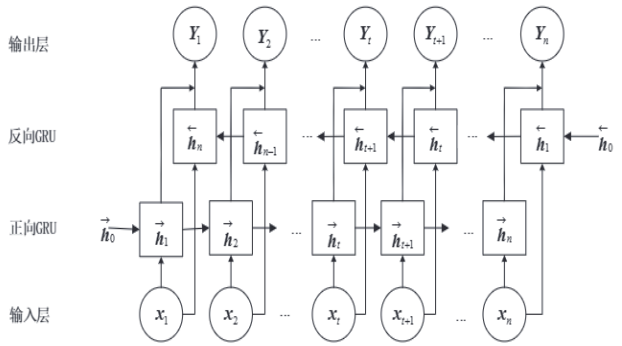


图 2 BiGRU 模型

BiGRU 通过正向 GRU 生成隐藏向量$\overrightarrow{\boldsymbol{h}}_t$捕捉从序列起始到当前位置的信息，反向 GRU 生成隐藏向量$\overleftarrow{\boldsymbol{h}}_t$捕捉从序列末尾到当前位置的信息，最终融合两个方向的信息获得上下文语义信息特征 $Y_t$。其中，每个 GRU 单元的结构如图 3 所示。

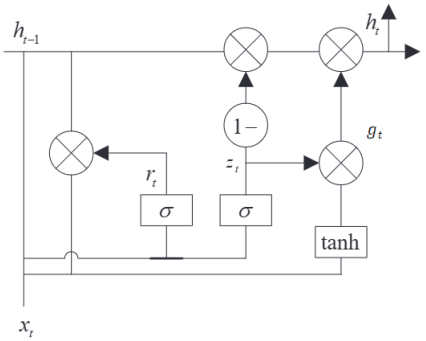


图 3 GRU 单元

GRU 模型的运行流程如下：

首先，根据上一时刻输入 $h_{t-1}$ 和当前时刻输入 $x_t$，计算重置门 $r_t$ 与更新门 $z_t$。重置门负责删除上一时刻输入中一些不必要的信息，更新门负责筛选出所需的新记忆以继续传递。

然后，根据经重置门处理后的上一时刻输入与当前时刻输入，计算出候选隐藏状态 $g_t$，生成新的信息。

最后，通过更新门和候选隐藏状态得出最终隐藏状态 $h_t$，作为下个时间步的输入。单个 GRU 模型的计算流程分别为：

$$z_t = \sigma(\boldsymbol{W}_z \cdot [h_{t-1}, x_t]) \tag{7}$$

$$r_t = \sigma(\boldsymbol{W}_r \cdot [h_{t-1}, x_t]) \tag{8}$$

$$g_t = \tanh(\boldsymbol{W} \cdot [r_t \odot h_{t-1}, x_t]) \tag{9}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot g_t \tag{10}$$

式中：$\boldsymbol{W}$ 为权重矩阵；$\sigma$ 为 sigmoid 函数；$x_t$ 为当前时刻输入；$h_{t-1}$ 为上一时刻输入；$g_t$ 为候选隐藏状态；$h_t$ 为最终隐藏状态。

1.4 全连接层

全连接层会将经 TextCNN 与 BiGRU 处理过的特征向量映射到最终输出，得到可读性评估级别，计算公式为：

$$y = \text{softmax}(\boldsymbol{W} \cdot \boldsymbol{x} + b) \tag{11}$$

式中：$y$ 表示每个类别的预测概率；$\boldsymbol{W}$ 表示权重矩阵；$\boldsymbol{x}$ 表示输入特征；$b$ 表示偏置项。

## 2 实验与分析

2.1 数据集

本文使用的数据集为整理后的 2010—2018 年的汉语水平考试 HSK 真题，按照 HSK 1~6 级划分为 6 个文本可读性等级，该数据集详细信息如表 1 所示。

表 1 实验数据统计

| 等级 | 训练集 | 验证集 | 测试集 | 总句数 | 总字数 |
|---|---|---|---|---|---|
| 1 | 587 | 73 | 54 | 1 173 | 10 078 |
| 2 | 890 | 103 | 109 | 1 861 | 25 698 |
| 3 | 1 036 | 146 | 128 | 2 948 | 53 303 |
| 4 | 763 | 102 | 106 | 2 810 | 82 787 |
| 5 | 932 | 106 | 125 | 8 279 | 273 631 |
| 6 | 368 | 42 | 51 | 7 220 | 267 642 |

2.2 参数设置

本实验在 PyTorch1.0.2 和 Python3.10 版本下进行，表 2 是实验超参数的设置。

表 2 参数设置

| 参数 | 说明 | 数值 |
|---|---|---|
| Batch __ size | 批量大小 | 8 |
| Embedding_size | 词向量维度 | 768 |
| $M$ | XLNet 层数 | 6 |
| Filter_size | 卷积核大小 | 2,3,4 |
| Num_filter | 卷积核数量 | 128 |
| Hidden_units | BiGRU 单元数 | 128 |
| Dropout | Dropout 保留比例 | 0.5 |
| Epoch | 训练周期 | 16 |
| $\mu$ | 学习率 | 5e-5 |

2.3 对比模型

为了检验本实验模型的性能，本文选择几个经典深度学习模型进行比较。

（1）TextCNN：以经过 Chinese-XLNet 训练后的词向量作为输入，利用 TextCNN 对特征进行提取，其中卷积核的数量和大小与本文设置一致，最终外接 softmax 进行分类。

（2）LSTM：LSTM 是一种特殊的递归神经网络，它擅长处理序列数据。以经过 Chinese-XLNet 训练后的词向量作为输入，利用多个 LSTM 网络进行特征提取，最终外接 softmax 进行分类。

（3）BIGRU：以经过 Chinese-XLNet 训练后的词向量作为输入，利用多个正向与反向的 GRU 网络进行特征提取，最终外接 softmax 进行分类。

（4）TextCNN-LSTM：以经过 Chinese-XLNet 训练后的词向量作为输入，利用 TextCNN 进行局部特征提取，之后再利用 LSTM 进行全局特征提取，最终外接 softmax 进行分类。

2.4 实验结果分析

本实验在汉语水平考试 HSK 真题数据集上将本文提出的 XLNet-TB 中文文本可读性评估模型与上述 4 种模型对比，实验结果如表 3 所示。

表 3 各模型各项性能对比

| 模型 | Acc | Precision | Recall | $F_1$ |
|---|---|---|---|---|
| TextCNN | 0.851 | 0.858 | 0.850 | 0.853 |
| LSTM | 0.841 | 0.844 | 0.840 | 0.841 |
| BiGRU | 0.867 | 0.868 | 0.865 | 0.867 |
| TextCNN-LSTM | 0.872 | 0.874 | 0.870 | 0.873 |
| XLNet-TB | 0.895 | 0.901 | 0.894 | 0.897 |

从表 3 可以看出，XLNet-TB 模型相较于其他模型的准确率等指标上均有提升，在汉语水平考试 HSK 真题数据集上的分类效果优于其他模型。XLNet-TB 模型利用 Chinese-XLNet 获取文本的动态词向量表示，并联合使用 TextCNN 与 BiGRU 获取文本局部、全局语义信息特征，增强了模型对文本语义信息的捕捉能力，从而取得了较好的文本可读性评估效果。

## 3 结语

针对现有模型在捕捉深层次语义信息方面的不足，本文提出了一种基于 XLNet-TB 分层网络架构的中文文本可读性评估模型，该模型通过 Chinese-XLNet 生成文本的动态词向量表示，利用 TextCNN 提取文本局部语义信息特征，再使用 BiGRU 通过正向与反向 GRU 提取上下文全局语义信息特征，两者结合增强了模型捕捉文本语义信息的能力，提升了对文本可读性评估的效果。在汉语水平考试 HSK 真题数据集上的准确率达到了 89.5%，优于其他模型。未来，将进一步考虑扩大数据集并在本模型的基础上融入更丰富的汉语语言特征，以不断优化模型的性能。

**参考文献：**

[1] 左虹，朱勇．中级欧美留学生汉语文本可读性公式研究 [J]. 世界汉语教学 ,2014,28(2):263-276.

[2] 刘苗苗，李燕，王欣萌，等．分级阅读初探：基于小学教材的汉语可读性公式研究 [J]. 语言文字应用 ,2021(2):116-126.

[3] 王蕾．初中级日韩学习者汉语文本可读性公式研究 [J]. 语言教学与研究 , 2017(5): 15-25.

[4] 于东，吴思远，耿朝阳，等．基于众包标注的语文教材句子难易度评估研究 [J]. 中文信息学报 , 2020,34(2):16-26.

[5] 杜月明，王亚敏，王蕾．汉语水平考试（HSK）阅读文本可读性自动评估研究 [J]. 语言文字应用 ,2022(3):73-86.

[6] 杨智渤．基于机器学习的汉语儿童阅读材料可读性评估方法研究 [J/OL]. 情报科学 ,1-16[2024-09-16].http://kns.cnki.net/kcms/detail/22.1264.G2.20240729.0916.002.html.

[7] 颜伟嘉．基于机器学习的 HSK 短文阅读测试文本可读性自动分析研究 [D]. 北京：中国石油大学 ,2022.

[8] 唐玉玲，张宇飞，于东．结合深度学习和语言难度特征的句子可读性计算方法 [J]. 中文信息学报 , 2022, 36(2): 29-39.

[9] ZULQARNAIN M, SAQLAIN M. Text readability evaluation in higher education using CNNs[J].Journal of industrial intelligence,2023,1(3):184-193.

[10] 沙政．基于深度学习的英文文本可读性度量研究 [D]. 重庆：重庆大学 ,2021.

[11] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: generalized autoregressive pretraining for language understanding[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems.NewYork: ACM, 2019: 5753-5763.

**【作者简介】**

倪佳成（1999—），男，江苏徐州人，硕士，研究方向：自然语言处理。

（收稿日期：2024-12-02）